FOLIA 232

Annales Universitatis Paedagogicae Cracoviensis

Studia Sociologica 9 (2017), vol. 1, s. 51–62
ISSN 2081-6642
DOI 10.24917/20816642.9.1.3

Filip Graliński, Daniel Dzienisiewicz, Piotr Wierzchoń
Uniwersytet im. Adama Mickiewicza w Poznaniu

U bram lingwistycznej szczęśliwości, czyli kulisy projektu Odkrywka: cyfrowe zasoby kultury jako źródło mas danych językowych

Streszczenie

Współcześnie coraz większa liczba materiałów historycznych, takich jak gazety, książki i dokumenty archiwalne, poddawana jest procesowi digitalizacji, a następnie udostępniana w formie cyfrowej w Internecie. Federacja Bibliotek Cyfrowych (FBC), gromadząca, przetwarzająca i udostępniająca informacje o internetowych zbiorach polskich instytucji nauki i kultury, liczy ponad 100 źródeł danych, w skład których wchodzą biblioteki cyfrowe, archiwa, repozytoria i inne. Spośród polskich bibliotek cyfrowych można wyodrębnić m.in. Wielkopolską Bibliotekę Cyfrową, Małopolską Bibliotekę Cyfrową, e-bibliotekę Uniwersytetu Warszawskiego i Jagiellońską Bibliotekę Cyfrową. Zbiory udostępniane przez te biblioteki, a także inne teksty elektroniczne są bogatym źródłem wiedzy o historii, kulturze, społeczeństwie i języku. Przedstawione źródła danych stanowią podstawę projektu Odkrywka, realizowanego przez pracowników Instytutu Językoznawstwa i Pracowni Systemów Informacyjnych Uniwersytetu im. Adama Mickiewicza w Poznaniu. Projekt ten ma na celu wykorzystanie bazy zdigitalizowanych tekstów w języku polskim do prowadzenia szybkich i efektywnych badań nad językiem, kulturą i historią Polski. W przedsięwzięciu w charakterze korpusu diachronicznego, liczącego setki tysięcy tekstów XIX i XX w., wykorzystywane są kolekcje upublicznione przez biblioteki cyfrowe oraz inne źródła internetowe. W artykule podane zostały najważniejsze wiadomości dotyczące projektu, zaprezentowano narzędzia wyszukiwania wyrazów i fraz oraz wykresy częstości. Poruszone zostało zagadnienie aktualnych badań oraz perspektywicznych analiz prowadzonych w oparciu o stworzony system.

Słowa kluczowe: digitalizacja, folklorystyka, językoznawstwo korpusowe, lingwochronologizacja, fotodokumentacja

Wprowadzenie. Rozwój bibliotek cyfrowych w Polsce

Powstanie polskich bibliotek cyfrowych spowodowane było m.in. serią kradzieży w bibliotekach tradycyjnych w końcu lat 90. XX wieku. W roku 1998 dokonano kradzieży w Bibliotece Polskiej Akademii Nauk (wydany w 1543 egzemplarz dzieła Mikołaja Kopernika *O obrotach sfer niebieskich*), a w 1999 w Bibliotece Jagiellońskiej (dzieła Galileusza, Keplera czy Bessariona). Wskutek tych wydarzeń podjęto pewne działania, które mają interesujące nas konsekwencje. Drogę do rozwoju polskich bibliotek cyfrowych otworzyło rozporządzenie Ministra Kultury i Sztuki z dnia

24 listopada 1998 r., które nakazywało bibliotekom szczególną ochronę zasobów, polegającą na:

1) sporządzeniu planu ochrony zasobów,
2) zabezpieczeniu przed zniszczeniem w schronach, budowlach ochronnych w obrębie jednostki organizacyjnej oraz zorganizowaniu odpowiedniej ochrony,
3) ograniczeniu udostępniania zasobu wyłącznie do celów naukowych i ekspozycyjnych innym bibliotekom lub instytucjom zapewniającym właściwe warunki ich zabezpieczenia,
4) utrwalaniu zasobu na innych nośnikach (http://isap.sejm.gov.pl/Details Servlet?id=WDU19981460955 [dostęp: 30.09.2016]).

Między innymi dzięki dyrektywie w punkcie 4. rozpoczęto w Polsce szeroką akcję digitalizacyjną. W roku 2002 powstała pierwsza biblioteka cyfrowa wykorzystująca oprogramowanie *dLibra* — Wielkopolska Biblioteka Cyfrowa. *dLibra* to program przeznaczony do gromadzenia, redagowania i udostępniania publikacji cyfrowych, opracowany w Poznańskim Centrum Superkomputerowo-Sieciowym (http://fbc.pionier.net.pl [dostęp: 30.09.2016]). *dLibra* wykorzystuje protokół OAI. Obecnie (2016) z oprogramowania *dLibra* korzysta ponad 100 bibliotek cyfrowych, gromadzących blisko cztery miliony publikacji cyfrowych.

Zbiory udostępniane przez biblioteki cyfrowe, a także inne źródła elektroniczne, legły u podstaw serwisu Odkrywka, projektu realizowanego przez pracowników Instytutu Językoznawstwa i Pracowni Systemów Informacyjnych Uniwersytetu im. Adama Mickiewicza w Poznaniu. Przedsięwzięcie wykorzystuje bazę zdigitalizowanych tekstów do prowadzenia szybkich i efektywnych badań nad językiem, kulturą i historią Polski. W projekcie w charakterze ogromnego korpusu diachronicznego wykorzystywane są setki tysięcy polskich tekstów XIX i XX w.

Geneza serwisu Odkrywka

Idea stworzenia serwisu umożliwiającego przeszukiwanie zdigitalizowanych tekstów historycznych początkowo realizowana była dwutorowo w celu zaspokojenia żywotnych potrzeb dwóch niezależnych dziedzin nauki, tj. językoznawstwa i folklorystyki.

Cezurą w dziedzinie wykorzystywania masowych kolekcji tekstów elektronicznych w badaniach lingwistycznych było ogłoszenie drukiem pracy *Fotodokumentacja*, *Chronologizacja*, *Emendacja: teoria i praktyka weryfikacji materiału leksykalnego w badaniach lingwistycznych* (Wierzchoń 2008b). Monografia zawiera wykładnię teoretycznych podstaw lingwochronologizacji — dziedziny lingwistyki zajmującej się przyporządkowaniem jednostkom języka datacji w postaci wartości rocznych (por. Smółkowa 1998). Celem poszukiwań lingwochronologizacyjnych jest odnalezienie najwcześniejszych świadectw użycia jednostek języka w korpusach tekstów historycznych. Wykorzystanie zdigitalizowanych tekstów zgromadzonych w bibliotekach cyfrowych w charakterze korpusu ma wiele niezaprzeczalnych zalet, spośród których można wymienić:

– łatwość wyszukiwania jednostek języka dzięki uprzedniemu poddaniu tekstów procesowi OCR,
– zadośćuczynienie potrzebie wierności cytatu.

Poprzez wierność cytatu rozumiane jest przedstawienie fragmentu tekstu w formie fotograficznej, tj. operacja fotodokumentacji. Fotodokumentacja jest nieodłącznym elementem teorii lingwochronologizacji (TLCH). Składają się na nią trzy stałe komponenty-fotografie:

– cytat,
– tytuł źródła,
– data publikacji.

Poniżej przedstawiono przykład fotodokumentacji dla najwcześniej odnalezionego użycia wyrazu *zakupoholik*:


Zakupoholik kupuje za dużo, rzeczy niepotrzebne, bez rozeznania, oddala się od rzeczywistości, nie umie się kontrolować, nie umie wybierać, popada w długi, wypisuje czeki bez pokrycia — charakteryzuje dziennik. Zakupy powodują szybkie, krótkotrwałe rozładowanie napięcia i ulgę. Uzależnieni zajmują się

SŁUPSK DZIŚ GŁOS CZYTELNIKÓW
GŁOS POMORZA
25 stycznia 1996 r.

Rycina 1. *zakupoholik*

Źródło fotografii: http://bibliotekacyfrowa.eu/Content/13268/07008-0001.djvu [dostęp: 30.09.2016]

Zagadnieniu lingwochronologizacji, zarówno w wymiarze teoretycznym, jak i praktycznym, poświęcono liczne prace naukowe (np. Wierzchoń 2008a, 2009, 2010).

Niezależnie od badań językoznawczych pojawiła się potrzeba stworzenia wyszukiwarki tekstów operującej słowami kluczowymi (tagami) (zob. Graliński 2013), która mogłaby wspomóc badania folklorystyczne. Planowano budowę serwisu, który początkowo miałby gromadzić wyłącznie teksty legend miejskich, w dalszej perspektywie mógłby jednak zostać poszerzony o inne tekstu folkloru. System miał być wyposażony w metawyszukiwarkę, tj. wyszukiwarkę zbierającą wyniki zwrócone przez inne wyszukiwarki oraz sortującą wyniki, np. poprzez zwracanie zapytań w porządku chronologicznym, usuwanie duplikatów. Docelowo narzędzie miało obejmować swym zasięgiem możliwie największą liczbę tekstów, począwszy od zdigitalizowanych tekstów historycznych (m.in. przez biblioteki cyfrowe), po teksty elektroniczne, które pierwotnie powstały już w środowisku cyfrowym. Planowano uwzględnienie m.in. stron internetowych, archiwów czasopism oraz tekstów dostępnych na forach internetowych. Dążono do automatyzacji wyszukiwania tekstów legendowych dzięki zastosowaniu uczenia maszynowego. Odpowiednia ilość danych tekstowych mogłaby bowiem stanowić dla systemu materiał uczący, co następnie umożliwiłoby mechaniczne rozpoznawanie wątków legendowych przez system.

System został zaprojektowany z myślą o badaczach-folklorystach, którzy nie byliby ograniczeni wyłącznie do korzystania ze zgromadzonych zasobów, ale mogliby także uzupełniać je zdobytymi we własnym zakresie tekstami. Mając na uwadze stopniowy rozwój projektu, postulowano także jego umiędzynarodowienie celem wzbogacenia serwisu o teksty folkloru zgromadzone przez światową społeczność folklorystów. Serwis Odkrywka stanowi ukoronowanie i odpowiedź na potrzeby obu nurtów badawczych.

Odkrywka od kuchni

Technicznie Odkrywka to system informacyjny stworzony na bazie silnika wyszukiwawczego Apache Solr (http://lucene.apache.org/solr [dostęp: 30.09.2016]). *Front-end* (część systemu bezpośrednio widoczna dla użytkownika) to aplikacja webowa oparta na środowisku Yesod (http://www.yesodweb.com [dostęp: 30.09.2016]).

Obecnie (wrzesień 2016) materiał tekstowy, który został zindeksowany (a zatem taki, który może być przeszukiwany), to:

- 3,2 mln publikacji (w wypadku periodyków „publikacja" to zazwyczaj pojedynczy numer, bywa jednak, że jest to cały rocznik czasopisma),
- 19,7 mln stron (zeskanowanych papierowych stron albo – w wypadku tych publikacji *born-digital*, dla których pojęcie strony nie wydaje się odpowiednie – całych tekstów),
- 18 mld wyrazów,
- 91 mld znaków.

Podczas indeksowania jako podstawową jednostkę tekstu („dokument" w terminologii systemów wyszukiwawczych) traktuje się stronę, tzn. wyszukiwarka jako „adres" wyniku wyszukiwania zwraca hiperodsyłacz i numer strony. Jest to pewne uproszczenie – w wypadku gazet lepiej byłoby zwracać pojedynczy artykuł, niestety stworzenie systemu „wycinającego" artykuły ze strony jest dużym wyzwaniem informatycznym (w praktyce odsyłanie do strony jest zazwyczaj wystarczające).

Jak podano wyżej, zebrany materiał to ponad 90 mld znaków. Obecnie jest możliwe tworzenie bardzo dużych korpusów „webowych" opartych na stronach WWW (Graliński, Borchmann, Wierzchoń 2016) jako przykład badań socjolingwistycznych czy wręcz socjologicznych, które można prowadzić dzięki takim korpusom, ale nawet w porównaniu z rozmiarem tego typu zasobów (rzędu kilkuset gigabajtów) objętość materiału tekstowego zindeksowanego w Odkrywce jest znaczna (szczególnie biorąc pod uwagę, że korpusy webowe obejmują tylko łatwo dostępny tekst współczesny).

Metadane

Odkrywka to wyszukiwarka pełnotekstowa (i to stanowi o jej sile w porównaniu z prostymi wyszukiwarkami bibliotecznymi operującymi wyłącznie na metadanych). Nie oznacza to, że metadane są nieobecne w Odkrywce – wręcz przeciwnie, w wyszukiwaniu można odwołać się do wielu pól metadanych. Omówimy tutaj trzy w praktyce najistotniejsze: tytuł, typ publikacji i czas.

Często dopiero kombinacja wyszukiwania pełnotekstowego i wyszukiwania w metadanych może prowadzić do interesujących nas informacji (np. „podaj wszystkie periodyki z sierpnia 1939 r., w których treści pojawia się słowo *mobilizacja*").

Tytuł

Pole tytułu jest indeksowane pełnotekstowo, tzn. można szukać tytułów zawierających zadane słowo.

Typ publikacji

W systemie Odkrywka wprowadzono typologię 52 klas publikacji. W tabeli 1 przedstawiono 12 najczęstszych typów publikacji. Jak widać, dominują periodyki (55,1% wszystkich publikacji), spora część publikacji została zaklasyfikowana jako „inne" (czy raczej ich typ jest nieznany – nie był dostępny w zewnętrznych źródłach i ręczne oznaczenie typu było zbyt czasochłonne). Książki stanowią jedynie 1,5% wszystkich publikacji, zazwyczaj są one jednak obszerniejsze (niż gazety), więc ich wkład do „masy" tekstowej nie jest aż tak mały.

Tabela 1. Typy publikacji

Typ	Odsetek publikacji
periodyk	55,1
inne/nieznane	37,5
książka	1,5
artykuł	1,2
fotografia	1,0
starodruk	0,7
pocztówka	0,6
manuskrypt	0,4
doktorat	0,3
mapa	0,3
druk	0,3
ulotka	0,2
pozostałe	1,0

Czas

Zdecydowanie najważniejszym polem metadanych jest pole czasu (utworzenia czy publikacji – dyskusja na temat tego rozróżnienia, czy właściwie jego braku, wymagałaby osobnego miejsca). O ile może brakować pozostałych pól metadanych, o tyle brak informacji o czasie czyni tekst praktycznie bezużytecznym, jako że pole czasu umożliwia:

- ograniczenie wyników wyszukiwania do zadanego okresu (np. dwudziestolecia międzywojennego),
- przedstawianie wyników w porządku chronologicznym,
- tworzenie wykresów częstości względem czasu.

Informacja czasowa może być zadana z różną dokładnością („rozdzielczością"). W tabeli 2 zestawiono publikacje dostępne w Odkrywce względem parametru dokładności. Jak widać, dla zdecydowanej większości publikacji znany jest rok, a dla ponad połowy (!) – nawet dzień publikacji.

Tabela 2. Dokładność informacji czasowej

Dokładność	Odsetek publikacji
dniowa	60,0
miesięczna	1,0
roczna	33,6
mniejsza niż rok	1,6
brak	3,8

Dokładność wyszukiwania większa niż roczna może okazać się przydatna, jeśli szukamy informacji dotyczących wydarzeń dokładnie umocowanych w czasie. Na przykład, aby odnaleźć reakcje prasowe na upadek tzw. meteorytu tunguskiego, możemy używać ogólnych słów kluczowych powiązanych z niewielkim, kilkudniowym „oknem" (30 czerwca 1908 plus kilka dni) (http://re-research.pl/pl/post/2016-09-03-00007-meteoryt-tunguski.html [dostęp: 30.09.2016]).

Porządkowanie wyników

Obecnie wyniki wyszukiwania można porządkować na trzy sposoby:

- według dopasowania wyników do zapytania (relewantności; upraszczając: im więcej razy elementy zapytania pojawiają się w tekście, tym wyżej tekst pojawi się w wynikach zapytania),
- chronologicznie – począwszy od najstarszych,
- chronologicznie – począwszy od najnowszych.

Może wydawać się to zaskakujące, ale domyślnie wyniki porządkowane są nie według relewantności, lecz chronologicznie. W praktyce taki wybór okazał się racjonalny – nawet jeśli nie poszukujemy po prostu najstarszego wystąpienia danego słowa, zazwyczaj im starszy tekst, tym w poszukiwaniach badawczych bardziej interesujący (bywa często, że nowszych wyników jest dużo, ale nie wnoszą one wiele do prowadzonych badań). Potrzeba korzystania z trzeciej opcji (odwrócony porządek chronologiczny) pojawia się bardzo rzadko.

Eksperymentalny charakter systemu Odkrywka

Należy zaznaczyć, że system Odkrywka ma ciągle charakter eksperymentalny i pilotażowy, w chwili obecnej jest „zamknięty", używany głównie przez jego twórców. W szczególności ulepszenia wymaga interfejs użytkownika, który raczej sprawiłby trudność nowym użytkownikom, gdyby otworzyć system dla szerszej publiczności.

Nie oznacza to, że Odkrywka w bieżącej wersji to system tylko demonstracyjny. Przeciwnie, w rękach wprawnego użytkownika Odkrywka, nawet w obecnym, niedoskonałym wydaniu, przynosi konkretne badawcze owoce, co wykażemy w dalszej części artykułu.

Możliwości systemu

W swej obecnej wersji system Odkrywka posiada interfejs, na który składają się trzy podstawowe narzędzia: wyszukiwarka, wykresy oraz dossier.

Wyszukiwarka

Jak wspomniano wyżej, podstawowym narzędziem w systemie jest wyszukiwarka. Domyślnie mechanizm zwraca wyniki dla określonych zapytań (wyrazów bądź fraz) w porządku chronologicznym wraz z liczbą wszystkich odnalezionych wyników. Przykładowo, dla zapytania *emancypantka* wyszukiwarka zwraca ponad 3500 wyników, por. fragment prezentujący 3 pierwsze rezultaty:


Search

Zapytanie emancypantka

Zacznij od wyniku o indeksie 0

Szukaj

Wyniki dla *emancypantka*:

Wszystkich wyników: 3501

1850 Pisma Klemensa Protasza. T.1 s. 99

96 iż znakomite Panie, najpieszezenil j wyelJOwane, pJ'ze cid 1loWOŻ)'tne lJles" alillY, Jolią lu'awie gwałtl m SW) ch kozaMw i poddmiezuMw, ciągną za hrody brol1at) ch i cu(łzisiejszych naszycb **emancypantek** wrzała krew oWYl'hIO ich praszczlll'c1, z czasów mnie mój

1850 Pisma polityczne s. 308

do sufizmów nicI)lko bardzo w)'szukanych, /lIc i bardzo niedcrookratyczn\'ch ucickać. hronić, co się jJronić nic (fa.. Prawda, Zł' P) I:mie , o prawach politycznych kobiet, albo na.igł6wnicjszydJ publicystów. Thiero; powstając w izhip dpp\Jtowall dJ przcciw wnios slowa: zupełna rrl'rczcnlllr

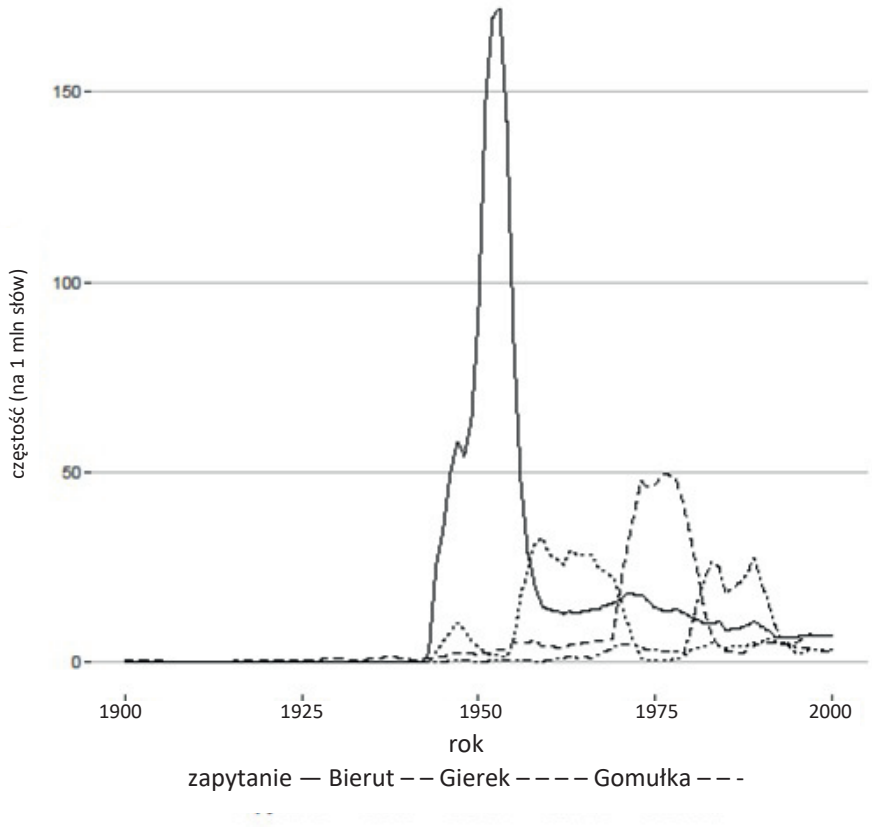
1852 Dziennik Warszawski Poświęcony Wiadomościom Krajowym i Zagranicznym, Lite

miernego wzrostu, kształtna, swobodna w ruchach, nie przymuszona w mowie, nieznajo przekonaniu, pod kategorję emancypacji. Lecz jakiej? to właśnie sęk! **Emancypantki**

Rycina 2. Wyniki dla zapytania *emancypantka*

Wykresy

System Odkrywka umożliwia kreślenie wykresów częstości występowania (słów bądź fraz) względem czasu (ściślej: roku). Wykres 1 to przykładowy wykres frekwencji nazwisk trzech I sekretarzy PZPR (jeden wykres może zawierać wyniki dla większej liczby wyrazów).

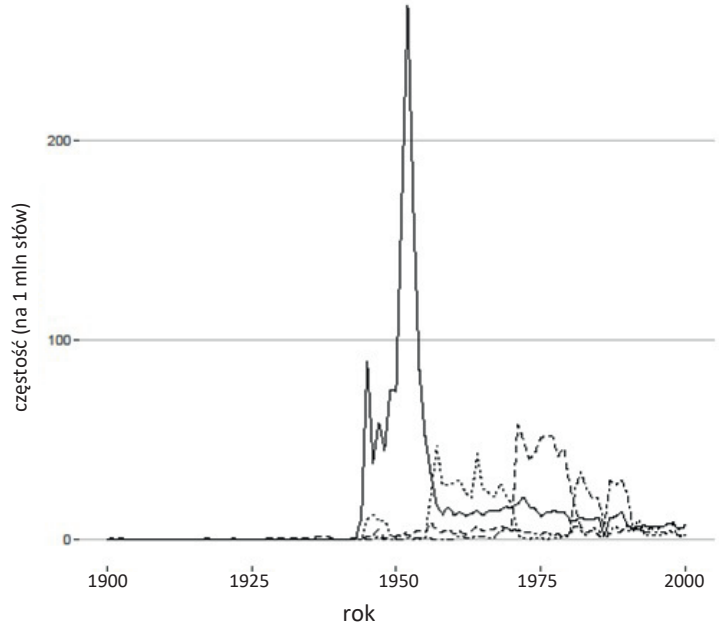


Wykres 1. Sekretarze (wygładzanie N = 3)

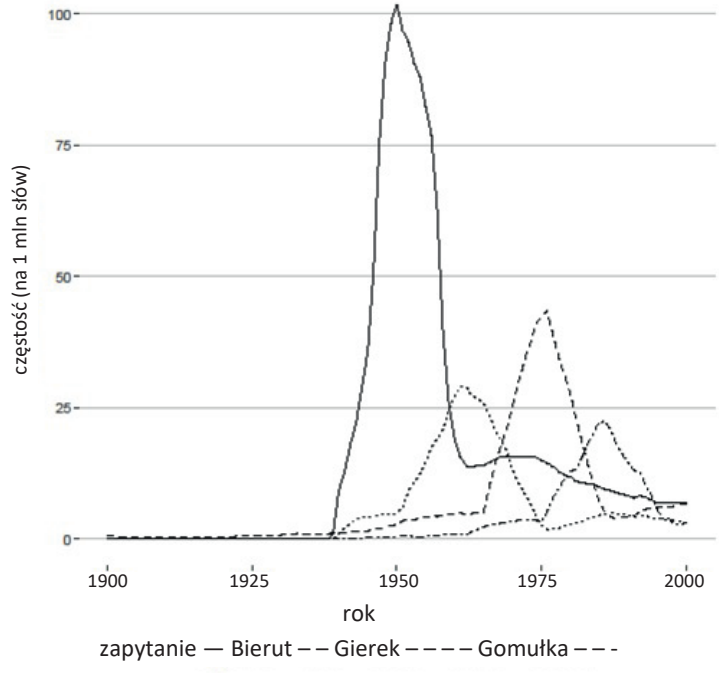
Pod względem tej funkcji Odkrywka przypomina system Ngram Viewer, który umożliwia generowanie tego typu wykresów dla kilku języków, ale nie dla polskiego (https://books.google.com/ngrams [dostęp: 30.09.2016]).

Wygładzanie. Istotnym parametrem jest tzw. wygładzanie. Domyślna wartość wygładzania to N = 3 (takie zostało też użyte na wykresie 1). Oznacza to, że tak naprawdę dla każdego roku na wykresie oddawana jest nie częstość występowania, lecz średnia z trzech lat — danego, poprzedzającego go oraz występującego po nim. Wygładzanie może wprowadzać w błąd, gdyż użytkownik nieświadomy tego efektu może odczytać, że w danym roku słowo pojawiło się (podczas gdy to tylko część „masy" wystąpień z roku kolejnego).

Dlaczego w takim razie stosuje się wygładzanie (w dodatku jako domyślną opcję)? Otóż pozwala ono abstrahować od przypadkowych jednorocznych fluktuacji i ułatwia dostrzeżenie istotnych trendów. Dla porównania podajemy tutaj wykresy bez wygładzania. Wykres jest jednak nadmiernie „postrzępiony" w porównaniu do wykresów poddanych wygładzaniu; wykres 2) i z większą wartością parametru wygładzania (N = 10, wykres 3).



Wykres 2. Brak wygładzania



Wykres 3. Wygładzanie N = 10

Dossier

System Odkrywka nie jest tylko „prostą" wyszukiwarką. Umożliwia on tworzenie dossier wycinków na zadany temat. Często bowiem dany temat badawczy wymaga użycia wielu zapytań, dla których tylko część rezultatów jest istotna. Funkcja ta ma ciągle charakter wysoce eksperymentalny. Nie poświęcimy jej tutaj więcej miejsca (choć jej efekty pojawią się w dalszej części artykułu).

Zastosowania badawcze

Serwis Odkrywka wychodzi naprzeciw potrzebom różnorodnych dziedzin nauki przyjmujących teksty za podstawę badań prowadzonych w ich obrębie. Od momentu powstania serwisu badania przeprowadzone w oparciu o narzędzia oferowane przez system Odkrywka były prezentowane na licznych konferencjach krajowych i międzynarodowych. Większość wyników badań nie została jednak jeszcze opublikowana. Zastosowanie narzędzi informatycznych w badaniach nad kulturą (o pojęciu tzw. *culturomics* zob. Michel i in. 2011) znacznie przyspiesza ich prowadzenie, formułowanie wstępnych hipotez oraz wniosków badawczych.

Dotychczas serwis Odkrywka wykorzystywany był m.in. do prowadzenia analiz dotyczących chronologii zapożyczeń japońskich (Dzienisiewicz, Wierzchoń 2016) oraz rosyjskich w języku polskim. Na podstawie zgromadzonych tekstów ustalono daty najwcześniejszych wystąpień japonizmów oraz rusycyzmów, takich jak np.: *bonza*, *gejsza*, *judoka* czy *bałagan*, *czystka*, *kołchoz*. W obu badaniach poruszony został także problem frekwencji form wariantywnych w różnych przedziałach czasowych (np. *judoka/dżudoka*, *pieriestrojka/pierestrojka/perestrojka*).

Ponadto przeprowadzono badanie nad kreatywnością językową Adolfa Nowaczyńskiego. Na podstawie materiału zawartego w *Słowniku bibliograficznym języka polskiego* (Wawrzyńczyk 2013) ustalono m.in., które wyrazy występujące w twórczości A. Nowaczyńskiego istniały w tekstach przed zastosowaniem ich przez pisarza oraz które jednostki stanowią indywidualizmy autora *Małpiego zwierciadła*. Podczas analiz wykorzystano rozszerzoną funkcję dossier, umożliwiającą stworzenie listy frekwencyjnej dla poszczególnych wyrazów w porządku dekadowym.

Podjęto również próbę opisu rozwoju słownictwa polskiego związanego z tematyką trucizn i trucia na przestrzeni XIX i XX wieku. Innym przykładem badania słownictwa tematycznego jest analiza najwcześniejszych świadectw wystąpienia nazw potraw typu *fast food* w polskich tekstach, a także znalezienie najwcześniejszych wzmianek dotyczących ich spożycia (http://re-research.pl/pl/post/2016-09-19-60013-hot-dog.html [dostęp: 30.09.2016]).

Kolejna praca porusza zagadnienie spreparowanych doniesień prasowych odnalezionych w prasie dwudziestolecia międzywojennego za pomocą narzędzi systemu Odkrywka. Porównanie relacji prasowych ze źródłami zagranicznymi, na które powoływali się autorzy artykułów, wykazało brak omawianych wiadomości w prasie francuskiej, angielskiej, meksykańskiej i amerykańskiej.

Narzędzia systemu Odkrywka zostały także wykorzystane do ustalenia frekwencji względnej form pierwszej osoby liczby pojedynczej w zależności od rodzaju podmiotu wypowiadającego się. Badanie to zostało szczegółowo opisane w artykule

'*He Said She Said' – A Male/Female Corpus of Polish* (Graliński, Borchmann, Wierzchoń 2016).

Zarysowane powyżej badania przeprowadzone przy wykorzystaniu systemu Odkrywka ukazują jego szeroką przydatność w badaniach tradycyjnie leżących w zakresie zainteresowania niezależnych od siebie dziedzin nauki.

Podsumowanie

Serwis Odkrywka pozwala na prowadzenie badań w zakresie m.in. takiej problematyki, jak: lingwochronologizacja, kreatywność językowa literatów i dziennikarzy, fluktuacje ortograficzne, legendy miejskie (szerzej: folklor), trendy społeczne czy historia (*fast food*). Lista dziedzin pozostaje jednak ciągle otwarta i badacz nie jest zmuszony do ograniczania się wyłącznie do powyższych ram. W przyszłości w oparciu o stworzony system planowane są również badania nad słowami-efemerydami, tj. wyrazami używanymi w języku wyłącznie w krótkich odstępach czasu. Ponadto interesującym zamierzeniem badawczym wydaje się być prześledzenie historii wybranych polemik prasowych, a także – szerzej – historii polskiej prasy XIX i XX w. Odkrywka mogłaby też być używana do wyszukiwania informacji biograficznych, nieujmowanych w dotychczasowych biografiach bądź biogramach, jak również, wraz z ciągłym przyrostem liczby tekstów, do poszukiwania informacji rozproszonych bądź zaginionych. Żmudne odnajdywanie zagubionych w tekstach okruchów historii pozwala bowiem na ich ponowne odkrywanie.

Bibliografia

Dzienisiewicz D., Wierzchoń P. (2016). *On the Japaneseness of Polish: A Linguochronological Approach*. Opuscula Iaponica & Slavica, t. 3, s. 53–76.

Graliński F. (2013). *Folklorystyka 2.0*. W: P. Grochowski (red.), *NETLOR. Wiedza cyfrowych tubylców*. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, s. 119–130.

Graliński F., Borchmann Ł., Wierzchoń P. (2016). *'He Said She Said' – A Male/Female Corpus of Polish*. W: N. Calzolari, K. Choukri, T. Declerck et al. (red.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).

Michel J.-B., Shen Y.K., Aiden A.P., Veres A., Gray M.K., The Google Books Team, Pickett J.P., Hoiberg D., Clacy D., Norvig P., Orwant J., Pinker S., Nowak M.A., Aiden E.L. (2011). *Quantitative Analysis of Culture Using Millions of Digitized Books*. Science 331(6014), s. 176–182.

Smółkowa T. (red.) (1998). *Nowe słownictwo polskie. Materiały z prasy lat 1985–1992*. Cz. I: A–O. Kraków: Instytut Języka Polskiego.

Wawrzyńczyk J. (2011). *Słownictwo nowopolskie. Redatacje*. Warszawa: Katedra Lingwistyki Formalnej UW.

Wawrzyńczyk J. (2013). *Słownik bibliograficzny języka polskiego*, t. 1–10. Warszawa: BEL Studio.

Wierzchoń P. (2008a). *Anti*. Poznań: Instytut Językoznawstwa Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Wierzchoń P. (2008b). *Fotodokumentacja. Chronologizacja. Emendacja. Teoria i praktyka weryfikacji materiału leksykalnego w badaniach lingwistycznych*. Poznań: Instytut Językoznawstwa Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Wierzchoń P. (2009). *Dlaczego fotodokumentacja? Dlaczego chronologizacja? Dlaczego emendacja? Instalacja gazowa, parking podziemny i „odległość niezerowa".* Poznań: Instytut Językoznawstwa Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Wierzchoń P. (2010). *Torując drogę teorii lingwochronologizacji*, Investigationes Linguisticae, t. XX, s. 105–185.

Strony internetowe

http://isap.sejm.gov.pl/DetailsServlet?id=WDU19981460955 [dostęp: 30.09.2016].

http://bibliotekacyfrowa.eu/Content/13268/07008-0001.djvu [dostęp: 30.09.2016].

http://re-research.pl/pl/post/2016-09-19-60013-hot-dog.html [dostęp: 30.09.2016].

http://re-research.pl/pl/post/2016-09-03-00007-meteoryt-tunguski.html [dostęp: 30.09.2016].

http://fbc.pionier.net.pl [dostęp: 30.09.2016].

http://lucene.apache.org/solr [dostęp: 30.09.2016].

http://www.yesodweb.com [dostęp: 30.09.2016].

https://books.google.com/ngrams [dostęp: 30.09.2016].

At the Gates of Linguistic Happiness, or the Backstage of the Odkrywka Project: Digital Resources as a Source of Language Data

Abstract

At present more and more historical materials such as newspapers, books and archival documents are being digitized and subsequently published on the Internet. The Federation of Digital Libraries which collects, processes and shares the information about the online collections of Polish cultural and scientific institutions contains over 100 sources of data. Digital libraries, archives, repositories and other sources can be found there. Among Polish digital libraries one can distinguish the Digital Library of Wielkopolska, the Digital Library of Małopolska, Warsaw University Digital Library and the Jagiellonian Digital Library. The collections which have been made available by these libraries are a rich source of historical, sociological, cultural and linguistic data. The above-presented sources of data are the basis of Odkrywka project. which is conducted by researchers from the Institute of Linguistics and the Laboratory of Information Systems at Adam Mickiewicz University in Poznań. The project's goal is to analyze Polish language, culture and history in an interdisciplinary way with the use of the gathered digitized texts. The collections made available by digital libraries are used as a large diachronic corpus with hundreds of thousands texts from the XIX and XX century. In the article the most important information about the project are given and the searching tools and automatically generated frequency graphs are presented. Moreover, past and current research and the prospects of development of the project are discussed.

Key words: digitization, folkloristics, corpus linguistics, linguochronologization, photodocumentation